Using Wordnet hierarchies to pinpoint differences in related texts

Ann Devitt
Computational Linguistics Group
Department of Computer Science
Trinity College Dublin
Ann.Devitt@cs.tcd.ie

Carl Vogel
Computational Linguistics Group & CCLS
Department of Computer Science
Trinity College Dublin
vogel@tcd.ie

Abstract

We present a means of comparing texts to highlight their informational differences. The system builds a Directed Acyclic Graph representation of the combined WordNet hypernym hierarchies of the nouns. Comparison of these yields a graph which distinguishes minor lexical & major content differences.

1 Introduction

As John Locke wrote in the seventeenth century

> No knowledge without discernment

In the 21st Century, the average human needs more than the average amount of discernment to get much knowledge from the vast profusion of information available to them. This paper addresses the question of distinguishing texts from each other. It is about similarities but also differences, detailing a structure by which differences of content may be distinguished, independent from surface differences.

The paper describes a system for modeling texts starting from its lexical items and supplementing these with "world knowledge" from WordNet and then comparing the content of these models. Section 2 discusses theoretical models of text which highlight the central role of lexis in text meaning and their use in Natural Language Processing applications. Section 3 details the representation built and the process of text comparison. Section 4 sets out the analysis for 3 pairs of texts. Finally there is a brief discussion of this work-in-progress and how it will progress.

2 Related Work

Much current research in the area of NLP is indebted to (Halliday and Hasan, 1976) for their exposition of the notion of cohesion as a combination of forces at work to knit a collection of disparate sentences into a text, a cohesive whole. They categorised the factors contributing to textual cohesion into grammatical devices—reference, substitution and ellipsis—conjunction and lexical cohesion.

In general terms, lexical cohesion relies on some similarity or relatedness among lexical items in a text. This assumption that the lexical items in a text are related systematically on more than just a grammatical level has fired research into means of text representation based on lexical items and also into the exploitation of such representations to tackle classic NLP problems.

Representations may be structured or unstructured. The analysis set out in (Halliday and Hasan, 1976) and developed in (Hasan, 1984) focuses on a structured representation: lexical chains of related items which through their interaction with each other make text coherent. (Hoey, 1991) further develops this idea as repetition nets which capture the relations between lexical items in a text. Many computational models take lexical chains as the premise for building representations of text, for example (Harabagiu and Moldovan, 1998).

Approaches such as Latent Semantic Analysis (Deerwester et al., 1990) do not attempt to build a structured representation of the relations between individual lexical items but give a global picture of text "meaning".

Given the availability of knowledge bases, computational approaches based on lexical cohesion can draw on more than just the lexical items to build a model of text. WordNet (Fellbaum, 1990) has been used extensively in this regard, see (Mihalcea and Moldovan, 1999), (Harabagiu, 1999), (Agirre and Rigau, 1996). The knowledge bases represented by large corpora are also used to supplement lexical information, systems include LSA (Deerwester et al., 1990), Vectile (Kaufmann, 1999).

These representations have been exploited to tackle many classic NLP problems such as word sense disambiguation (Mihalcea and Moldovan, 1999), meronymy resolution (Markert and Hahn, 2002), pronominal resolution (Harabagiu and Maiorano, 1999), topic identification and text segmentation (Hearst, 1997), (Kaufmann, 1999), textual inference (Harabagiu and Moldovan, 1998).

This paper follows in this vein, on the premise that just the collection of lexical items have much to tell about the structure and content of a text without taking other structural aspects into account. It presents a very simple representation of lexical items in a text enriched with WordNet hypernym relations which is used to highlight types of differences between texts. The tasks of text categorization and information retrieval deal with gauging similarities among texts: texts and corpora or texts and query strings where a high measure of similarity denoting success. Shifting the focus to text differences may yield interesting results for relevance feedback where a measure of similarity has been calculated and categorisation of where individual texts differ from some desired target could be useful.

3 Approach

The approach used here is quick and dirty. The aim is to build some representation of a text, in this instance news text, then to compare representations of related stories to identify their differences.

The representation is built using just nouns. This is clearly a short-coming, however, for a quick and dirty approach, it suffices for the moment.

In order to extract the nouns, the texts are first tagged using LT POS, a tagger and chunker from the Language and Technology Group at the University of Edinburgh. The hypernym hierarchy for all senses of each noun is then extracted from WordNet. These synset lists are then merged into a directed acyclic graph to represent the text, ambiguities and all. These graphs form the basis for a comparison of texts.

3.1 The Representation

The aim is to produce a useful representation which could be generated from free text, with minimal pre-processing and no disambiguation.

WordNet provides a rich knowledge base for English in which concepts, termed synsets or synonymy sets, are linked by semantic relations. The system detailed here uses the WordNet hypernym "IS-A" relation. For each lexical item in the text, the hypernym hierarchy from beginning to end is added to the representation of the whole. The final structure is a directed acyclic graph containing a subset of the WordNet hypernym linked structure. Each node in the graph represents a WordNet synset and a count of its frequency in the text. The intuition is that the main entities and topic areas are those linked by frequently traversed edges in the graph. This structure captures the lexical chains formed by reiteration—through general nouns, synonymy and superordinates or just repetition—as edges that have been traversed more than once.

For example, *Text1*, the following short text, based on an extract from (Halliday and Hasan, 1976) p. 279.

> There's a boy climbing that tree. The idiot will fall if he's not careful. Elms are not very sturdy. The poor child might hurt himself.

yields, among other chains, these lexical chains connecting boy-child-idiot and tree-elm.

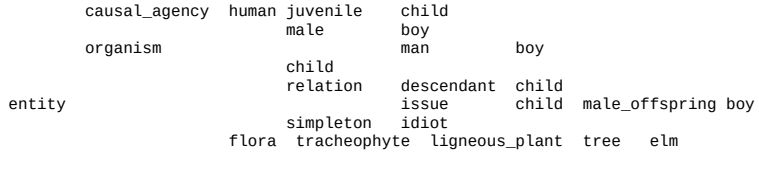

Figure 1. Partial hypernym graph for Text1

The text graph also contains spurious paths such as:

```
abstraction    attribute    form    figure    plane_figure    tree
```

Figure 2. Path from hypernym graph for Text1

While the flora-tree path is reinforced by the addition of the hypernym path for elms, the abstraction-tree path is not reinforced by connections with any other nodes and indeed remains the only node stemming from the root abstraction in the text.

In practice, graphs, such as figure 1, are stored as adjacency matrices of synset identifiers to facilitate comparison with other texts.

3.2 Comparison

The comparison of the matrix representations of two texts, *T1* and *T2*, yields another matrix of synset nodes. The comparisons and resulting matrices used in this approach are the following:

- **T1**$\setminus$**T2**: the synset nodes present in T1 but not in T2
- **T2**$\setminus$**T1**: the synset nodes present in T2 but not in T1
- **T1**$\cap$**T2**: the intersection of the two texts, all synset nodes present in both texts

If two texts are on completely different subjects, the difference matrices, *T1*$\setminus$*T2* and *T2*$\setminus$*T1*, contain most of the original matrices, T1 and T2, respectively. If they are quite similar, the difference matrix will contain a very restricted subset of the input synset nodes and the similarity matrix contains the lion's share. New entities are represented by complete paths in the matrix, other lexical choices may appear as absence or presence of a few additional nodes. An empty or almost empty difference matrix can indicate two things: the texts are very similar or one text subsumes another.

If the intersection of two texts is significant, a comparison of the *T1*$\setminus$*T2* and *T2*$\setminus$*T1* matrices can give an indication of which of the two texts is more informative on a particular subject, i.e. whether they are sufficiently different to warrant reading both or whether one subsumes the other. Possible relevance feedback comments include "if you've read this, you probably don't need to read that" or "read one or the other but don't worry about both" or "whatever they're saying, it's something very different, read both".

The statistical test used to determine significance of the results of comparison is chi-square. (Kilgarriff, To appear) states this performs best among other tests evaluating similarity within and across corpora and this is a related task.

Having computed T1$\cap$T2 T1$\setminus$T2 and T2$\setminus$T1, we formulate a chi-square contingency table of the form:

	Shared Text & One Part	**The Other Part**
T1	T1$\cap$T2 + T1$\setminus$T2	T2$\setminus$T1
T2	T1$\cap$T2 + T2$\setminus$T1	T1$\setminus$T2

Table 1: Chi-square contingency table schema

Thus, the rows in each case sum to the total text size (words, in the case of word based comparison, and total nodes for the two texts for the proposed WordNet derived comparison). The chi-square contingency table is used to test whether the null hypothesis, that there is no impact of the influence of text choice (the row) on the distribution in the column, may be rejected. Statistical significance means that the hypothesis may be rejected as an asymmetry exists in the contribution of one of the two texts to the overall total. Lack of significance indicates that the texts are very dissimilar to the point of not being informationally comparable. In other words, each row is bounded by (words or synsets) the total set of types derived from the union of the texts, but the columns isolate differential contributions of the text pairs. Thus, significance in the Chi-square test is directly related to differences in the individual contributions of the texts, albeit mitigated by the contribution of the $T1 \cap T2$ term which is a frequency count of their joint contribution.

The method allows pairwise comparisons of individual texts, and requires round-robin comparison to identify overall uniqueness in information contribution from a larger body of texts. Although examples are not included in the abstract, the very programs that allow computation of textual and WordNet node overlap also admit indexing of the documents to supply exact 'pinpointing' of putative informational differences.

The Chi-Square test as outlined bounds the entire comparison by the cardinality of synsets invoked by the two compared texts together. The row effect is the differential effect of the unique contribution of each individual text on that total cardinality. The column effect is the total contribution of each individual text in an algebraically constructed isolation.

Suppose a larger T1 ∪ T2 than T1 ∩ T2. Then, even if there is a size difference in T1 and T2, it is possible to measure the larger contributor to T1 ∪ T2, and hence measure information difference. If T1 ∪ T2 is comparable to T1 ∩ T2, then T1\T2 is comparable to T2\T1. If there is a disparity in T1 ∪ T2 compared with T1 ∩ T2, then it is possible to compare the relative contribution of T1 and T2, should the difference result from sample size. However, in the texts we experiment with, sample sizes are comparable.

The contingency table as constructed is bounded by the sizes of the overall set of unique concepts, that of the individual contributions, and that which is common to both.

The next section evaluates the system for 3 pairs of real news texts.

4 The Texts

This section provides examples from a corpus of news texts of how the system described in this paper can distinguish the degree of difference between texts. The texts, printed in Appendix A, are titled Arsenic1, Arsenic2, Bomb2, Shooting and Budget. The results discussed below are based on an evaluation of the comparison matrices using the chi-square significance test discussed above, with a significance level of p≤0.05, the chi-square value should be ≥ 3.84.

4.1 Same topic: Arsenic1 and Arsenic2

The two texts, Arsenic1 and Arsenic2, relate the same story of a mass poisoning at a Church in Maine and the death of the main suspect. However, coming as they do from different sources, the story is told somewhat differently in each.

Table 2 details the results of a comparison of the adjacency matrices of the two texts. The chi-square value, 9.32, indicates that the distribution of nodes is statistically significant, that the texts are related, and that their individual contributions beyond the intersection is interesting.

Chi-square	**9.32 ($p \leq .01$)**		
Measure	**T1∩T2**	**T1\T2**	**T2\T1**
Nodes	436	314	381
Percentage	39%	28%	33%

Table 2: Arsenic1 *vs* Arsenic2 node comparison

A comparison of the texts based purely on the lexical items in the texts yields the results set out in table 3. For significance at the .05 level at 1 df, chi-square should be greater than or equal to 3.84. Here, chi-square is 0 indicating that the word distribution is not significant, that the texts are not related. This points to an aspect of the intended contribution of our work: given the small sample sizes, text-based (that is, explicit word or lemma based) measures of informational contribution of two texts suffers from data-sparseness. While it is clear that POS tagged comparisons suffer from excess noise, our intention is to use the intermediate level semantic tagging supplied by activated WordNet nodes. This intermediate representation on these two short texts that are clearly related, but which also clearly involve substantially distinct vocabularies, provides an initial indication that the approach is viable.

Chi-square	**0 ($p \leq 1$)**		
Measure	**T1∩T2**	**T1\T2**	**T2\T1**
Words	96	108	108
Percentage	30%	35%	35%

Table 3: Arsenic1 *vs* Arsenic2 **word** comparison

4.2 Unrelated topics: Bomb2 and Budget

The last section demonstrated that our approach can generate useful information: "these two articles are related and both contribute distinct bits of information, so probably you should read both, with an eye to what the index mechanism flags". A contrasting statistic is also necessary if one is to receive advice of the form: "while noodle.news has classified these two articles as about the same topic, they actually contribute identically to their sum information" (ie. either they are identical or utterly unrelated).

The chi-square value in table 4 indicate that any similarities between texts Bomb2 and Budget are not statistically significant. For significance at the 0.05 level, the chi-square value should be ≥ 3.84.

Chi-square	**0.0020 ($p \leq 1$)**		
Measure	**T1∩T2**	**T1\T2**	**T1\T2**
Nodes	146	407	406
Percentage	15%	42.5%	42.5%

Table 4: Bomb2 *vs* Budget node comparison

Closer analysis of the intersection matrix, Bomb2∩Budget, yields an outline of what kind of similarities exist between the two texts. A large portion of the matrix contains paths ending on generic terms, such as:

skilled worker
electrical device
cognitive process

While the difference matrices contain the more specific terminal nodes of these paths.

- Bomb2: skilled worker to man / serviceman
- Budget: skilled worker to minister

This would suggest that all texts have a baseline similarity—being for the most part about entities, etc. An extended system should take into account the shape and nature of the intersection matrix, as well as its size, to determine whether the similarities are baseline or significant similarities.

4.3 Related topics: Bomb2 and Shooting

As a critical evaluation of our idea, we examine an intermediate comparison. Bomb2 and Shooting both recount a story of a shooting in Belfast. The events themselves are not related and happened at separate times. The subject matter, however, is similar. Again, the chi-square value in Table 5 are significant, indicating that the texts are related.

In this instance, the word comparison data also produces a significant result, with a chi-square value of 32.9, p value ≤ 0.001.

This reveals that our notion of topic tracking is not refined enough to base comparisons on texts automatically identified to be about the same *events*. However, it correspondingly demonstrates success in identifying informationally distinct articles about related *event types*: someone interested in one of the events may well be interested in the other on the basis of their common thematic content.

Chi-square	**235.30 ($p \leq .001$)**		
Measure	**T1∩T2**	**T1\T2**	**T2\T1**
Nodes	161	392	119
Percentage	24%	58%	18%

Table 5: Bomb2 *vs* Shooting node comparison

5 Discussion

Conclusive results would require an analysis of far more data than has been presented here. This task is currently in progress.

However, these preliminary analyses would suggest that this approach can discriminate differences in texts at least better than a baseline word comparison approach. The system outputs a quite dependable measure of similarity or differences between texts and also what these differences or similarities are—the terms associated with the synset.

6 Future Work

Our proposal is very much a work in progress, the aim to develop the representation to provide a basis for comparison of text for many criteria—content, difficulty, genre, etc. Some specific areas for development include: **Parts of speech**: The motivation for excluding all parts of speech besides nouns was to have a working model on which to begin experiments. The ultimate aim, however, is to incorporate all parts-of-speech in the text representation. **WordNet relations**: WordNet provides much more varied resources than just the hypernym hierarchy among nouns. The other relations: antonymy, meronymy and holonymy should also be exploited, as in, for example, (Harabagiu, 1999). **Dimension reduction or path-finding**: The inclusion of all parts of speech and of more WordNet relations entails an explosion in the size of the matrix bringing attendant of issues of how to deal with such a large data structure. **Ambiguity**: The DAG is not disambiguated so every ambiguous word introduces or strengthens spurious

nodes and edges. This aspect can become a task in itself—a hypernym hierarchy similar to that described here has already been used to a certain success for ambiguity resolution (Agirre and Rigau, 1996). It could also be exploited to some extent. There are cases when the level of ambiguity of a text may be a significant feature, for example, in deciding on suitability of a text for a particular readership. **DAG structural analysis**: As noted in section 4.3, the shape, as opposed to the contents, of the graph produced for an individual text is a telling characteristic. A comparison of structures across genre could yield interesting results.

References

Eneko Agirre and German Rigau. 1996. Word sense disambiguation using conceptual density. In *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen.

Association for Computational Linguistics. 1999. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, University of Maryland.

S Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407.

C. Fellbaum. 1990. *WordNet,an electronic lexical database*. The MIT Press.

J Flood, editor. 1984. *Understanding reading comprehension*. International Reading Association, Delaware.

Michael A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman.

Sanda M Harabagiu and Steven J Maiorano. 1999. Knowledge-lean coreference resolution and its relation to textual cohesion and coherence. In HarabagiuMaiorano99 (Ass, 1999), pages 29–38. Workshop on the relation of discourse/dialogue structure and reference.

Sanda M Harabagiu and Dan Moldovan. 1998. A parallel system for text inference using marker propogations. *IEEE Transactions in Parallel and Distributed Systems*, pages 729–747.

Sanda M Harabagiu. 1999. From lexical cohesion to textual coherence: - a data driven perspective. *Journal of Pattern Recognition and Artificial Intelligence*, 13(2):247–265.

Ruqaiya Hasan. 1984. Coherence and cohesive harmony. In Flood (Flood, 1984), pages 181–219.

Marti Hearst. 1997. Text-tiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Michael Hoey. 1991. *Patterns of Lexis in text*. Oxford University Press.

Stefan Kaufmann. 1999. Cohesion and collocation: Using context vectors in text segmentation. In Kaufmann99 (Ass, 1999), pages 591–595.

Adam Kilgarriff. To appear. Comparing corpora. *International Journal of Corpus Linguistics*.

John Locke. 1964. *An essay concerning human understanding*. Collins, 5 edition.

Katja Markert and Udo Hahn. 2002. Understanding metonymies in discourse. *Artificial Intelligence*, 135:145–198.

Rada Mihalcea and Dan Moldovan. 1999. A method for word sense disambiguation of unrestricted text. In *Proceedings of ACL '99*, pages 152–158, Maryland, NY, June.

A Appendix: News Texts

The following news texts are from RTE Interactive, the Irish National Broadcaster's internet division and from the Google News site

A.1 Arsenic1[1]

Maine Police Link Dead Man to Arsenic Case Sat May 3, 2003 01:32 PM *CARIBOU, Maine (Reuters)—Maine police on Saturday linked a man who died of a possibly self-inflicted gunshot wound to the arsenic-tainted coffee poisoning at a local church that killed one parishioner and sickened at least 15 others.*

State Police Col. Michael Sperry told reporters that Daniel Bondeson, 53, who died on Friday evening at a hospital in Caribou, was involved in the poisoning at the church that has rocked New Sweden, Maine, a town of about 600 people near the Canadian border.

"We have linked the shooting to the death at the church," Sperry said, explaining that police had found key information at Bondeson's farm house where he was found shot. Bondeson died after emergency surgery.

Sperry however declined to describe the information that police had found, and declined to say whether police had discovered a suicide note or arsenic. A search of Bondeson's Woodland, Maine home was expected to take several days.

"This was reported as a self-inflicted gunshot wound to us," Sperry said, adding however that police won't say for sure what happened until the body is autopsied on Monday.

The mystery began last Sunday when parishioners at the Gustaf Adolph Lutheran Church in New Sweden suddenly fell ill after services.

[1]Text via Google News, 3 May 2003

One man, 78-year-old Walter Morrill, died after drinking the coffee and several others were hospitalized. Bondeson, a bachelor, was a member of the church but did not attend last week's sermon. However his brother and sister were there but did not drink anything, parishioners said.

Police earlier this week said the concentration of arsenic led them to believe it had been put into the coffee deliberately. They called in the Federal Bureau of Investigation to gather fingerprints and DNA samples from as many as 50 parishioners to try and find the killer.

Arsenic can kill quickly if consumed in large quantities, although small, long-term exposure can lead to a much slower death. It canm give a strong bitter taste to food or beverages it contaminates.

A.2 Arsenic2[2]

Police: Shooting may be tied to poisoning. 5/3/2003 1:06 PM *NEW SWEDEN, Maine (UPI)—Investigators in the small northern Maine town of New Sweden searched Saturday for possible links between a fatal shooting and a church arsenic poisoning case.*

Daniel Bondeson, 53, died Friday evening, shortly after he was found shot at his home in Woodland, adjacent to New Sweden. Police said Bondeson was not a suspect in the church poisoning, but searched his home for any connection to the laced church coffee that killed one elderly man and sickened at least 15 other people. Two of the victims were still in critical condition.

"We won't make a determination until the autopsy (on Bondeson's body) Monday," Col. Michael Sperry, chief of the Maine State Police, said Saturday outside the Caribou courthouse.

"This is an open investigation, and we are still looking at who is involved," he said. Investigators are treating the arsenic case as a homicide.

"The FBI is still very much involved," adding the arsenic poisoning has had "a huge impact on this small community," about 20 miles south of the Canadian border.

The poisoning occurred Sunday after services at the Gustaf Adolph Lutheran Church in New Sweden, a community of about 620 in the potato-farming region in rural northern Maine.

Among the two dozen parishioners who drank the coffee was Walter Reid Morrill, 78, who died Monday. Morrill was a member who lived next to the church and served as caretaker and head usher.

The remaining coffee in the percolator contained "high levels" of arsenic, said Stephen McCausland, spokesman for the Maine Department of Public Safety.

He said tests confirmed the arsenic was not in the unbrewed coffee, in the tap water or in the sugar.

Word that someone apparently had deliberately put arsenic in the coffee frightened area residents.

"This is a small community where everyone knows each other," McCausland told United Press International. "We don't know whether (the perpetrator) was among them or someone from outside, but the focus of our investigation now is to find who is responsible for introducing the arsenic, and why."

[2]Text via Google News, 3 May 2003

A.3 Bomb2[3]

Man due in court for bomb attempt

A 34-year-old man is due in court in the North charged in connection with an attempted firebomb attack in Belfast city centre at the weekend.

The man has been charged with possession of explosives with intent to endanger life and conspiracy to cause an explosion.

He was arrested in a major security operation in the city centre last Sunday night, during which a second man was shot twice by police.

Police discovered a device consisting of two gas cylinders and two pipe bombs linked to a number of containers of flammable liquid in a car abandoned outside the motor tax office at Upper Queen Street at the weekend.

A timer device had been attached and activated.

The shooting of the second man by police is being investigated by the North's Police Ombudsman, Nuala O'Loan.

A.4 Shooting[4]

8SHOT. R1 SM 05-12-2002 07.51 *Two men have been injured in paramilitary style attacks in Ardoyne in north Belfast. A 27 year old man was shot in the leg; he was found near Ardoyne Avenue shortly after eight o'clock last night. And a 20 year-old man was also shot in the leg in the garden of a house at Butler Walk in the city. Police said the incident happened at around midnight.*

A.5 Budget[5]

Minister briefs Cabinet on Budget. 041202TM12.50

The Government's Budget for 2003 will be unveiled in the Dil this afternoon by the Minister for Finance, Charlie McCreevy. It is expected that only very small tax reductions will be included, while increases in Social Welfare payments will generally be kept in line with the rate of inflation.

The Minister is expected to make a budgetary provision to pay the 25% backdated element of the Public Sector Benchmarking pay awards. A freeze on public sector numbers as well as a three-year programme to reduce staff levels in the public sector is also expected be announced. Mr McCreevy briefed the Cabinet at an early morning meeting ahead of this afternoon's Dil debate. Coverage of the Dil debate will be broadcast during a special 5-7 Live programme on RTE Radio One from 3.30pm and on Raidio na Gaeltachta from 4.08pm. Special Budget programmes will also be broadcast on RTE One Television from 3.35pm, on Network Two from 4.30pm and on TG4 at 8.30pm this evening during which there will be a phone-in with a panel of experts. RTE Radio One, 2FM, Raidio na Gaeltachta and Lyric news bulletins will provide reports of Budget announcements. And following the 9 o'clock news on RTE One Television, there will be further interviews, analysis and discussion about the Budget. The RTE website and Aertel will also provide up-to-the minute reports.

[3]Text from RTE Interactive, 27 Nov 2002

[4]Text from RTE Interactive, 5 Dec 2002

[5]Text from RTE Interactive, 4 Dec 2002